CARNAC-LR: De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets

Camille Marchet, Lolita Lecompte, Corinne da Silva, Corinne Cruaud, Jean-Marc Aury, Jacques Nicolas, Pierre Peterlongo

▶ **To cite this version:**

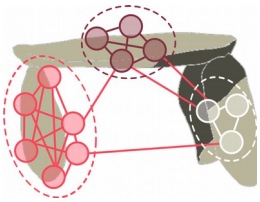
Camille Marchet, Lolita Lecompte, Corinne da Silva, Corinne Cruaud, Jean-Marc Aury, et al.. CARNAC-LR: De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets. RECOMB-seq 2018 - Eighth RECOMB Satellite Workshop on Massively Parallel Sequencing, Apr 2018, Paris, France. pp.1-2. hal-01929963

HAL Id: hal-01929963
https://hal.science/hal-01929963

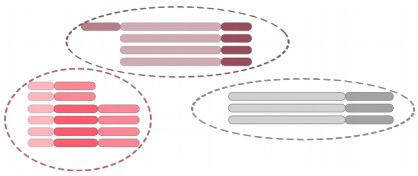
Submitted on 21 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



CARNAC-LR:
De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets



Camille Marchet[1], Lolita Lecompte[1], Corinne Da Silva[2], Corinne Cruaud[2], Jean-Marc Aury[2], Jacques Nicolas[1], Pierre Peterlongo[1]

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

[2]Commissariat à l'Energie Atomique (CEA), Institut de Biologie Francois Jacob, Genoscope, 91000, Evry, France

Goal: *de novo* cluster Nanopore reads per expressed genes

Data: Nanopore 1D reads from mouse transcriptome sequenced with MinION (accession number: ERP107503)

Results:

★ State of the art does not perform well on ONT reads

★ We introduce CARNAC-LR, a new clustering approach designed for long reads

★ Validations on mouse transcriptome

Benchmark 1: community detection algorithms

	Recall	Precision	Jaccard index
Single link	76%	<15%	<0.1
Louvain	89%		
Modularity	61%	<75%	<0.5
CPM	79%		
CARNAC-LR	65%	98%	0.79

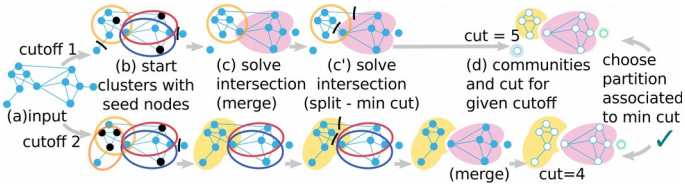
Benchmark 2: sequence clustering tools

	Recall	Precision	Status
Starcode	NA	NA	error
Tofu	NA	NA	not applicable
SEED	0	0	run
CD-HIT	27%	99%	run
CARNAC-LR	65%	98%	run

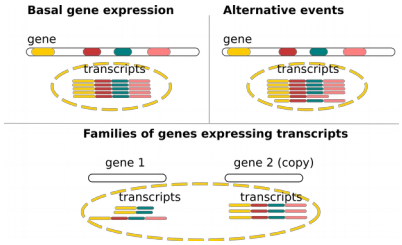
Algorithm overview:

Key ideas:

- maximize local edge density
- minimize cut size
- partition the graph

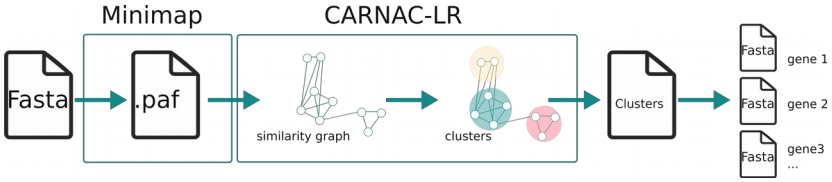


Expected clusters:



Pipeline overview:

From reads to clusters per expressed gene

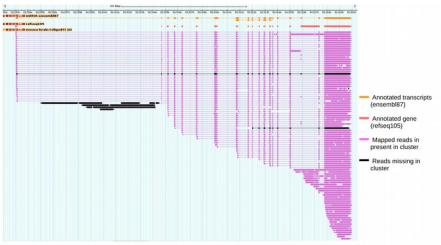


★ C++11 and Python 3

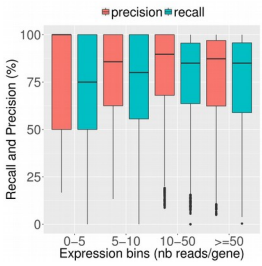
★ GPL license

★ **available on Github**

Results on whole mouse transcriptome:



Output graphical example for mouse Picp5 gene



Clusters **purity** and **completeness** assessed using mapping strategy (BLAT+est2genome)

Performances:

For 1 million reads

→wallclock 3 hours (40 threads)

→memory: 30G

Work in progress:

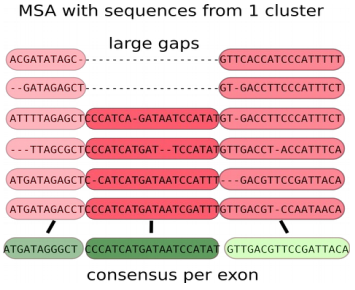
Goals:

-Identify alternative isoforms from CARNAC-LR's clusters

-Propose one consensus per isoform

Key ideas:

- intra-cluster multiple sequence alignment
- detect alternative blocks (exons)
- separated block consensus computation



Acknowledgments

★ Genoscope platform

★ ASTER ANR

★ Genouest team & infrastructures

Main achievements

★ Clusters *de novo* ONT reads by expressed genes

★ Scales a whole mouse transcriptome

★ Performs better than state of the art on ONT reads

★ Validated using comparison to mapping strategy on real data

Tool:

github.com/kamimrcht/CARNAC-LR

Preprint:

biorxiv.org/content/early/2018/03/26/170035

Contact : camille.marchet@irisa.fr